An Automatic Finite-Sample Robustness Metric for Bayes & Beyond: Can Dropping a Little Data Change Conclusions?

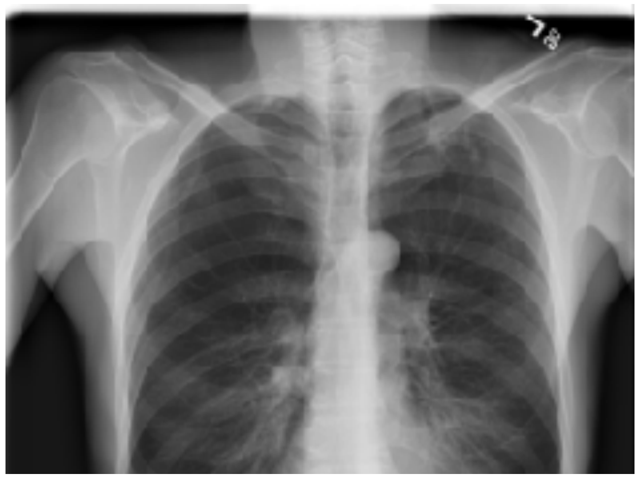
Speaking today:
Tamara Broderick, Ryan Giordano (MIT)

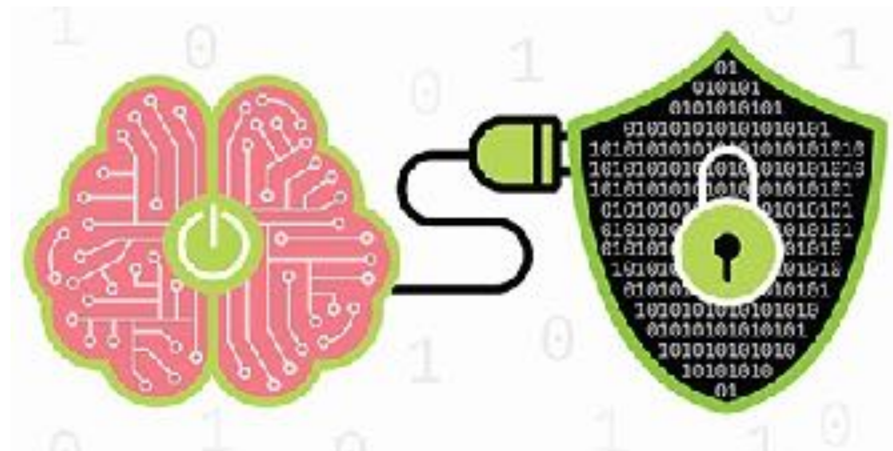
Work with:
Rachael Meager (LSE)

When can I trust my data analysis?

When can I trust my data analysis?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions

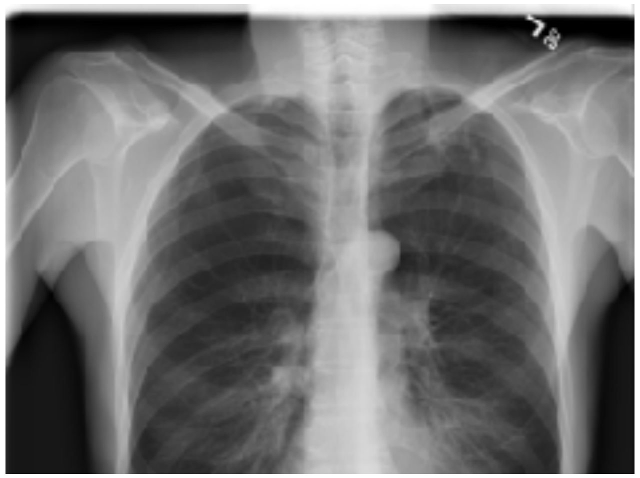


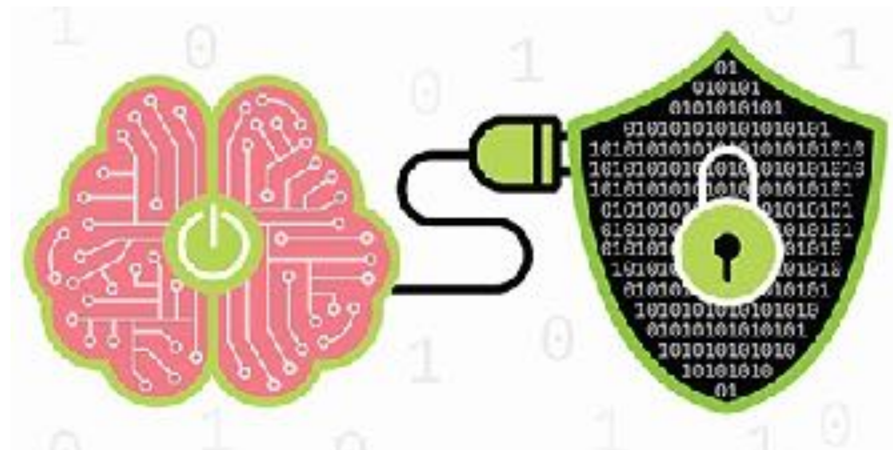


[https://stanfordmlgroup.github.io/projects/chexnet/, https://www.ibmbigdatahub.com/blog/cyber-security-powered-ai-and-machine-learning, https://www.fairwinds.org/inside-fairwinds/fraud-protection-center/report-fraud.html]

When can I trust my data analysis?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions



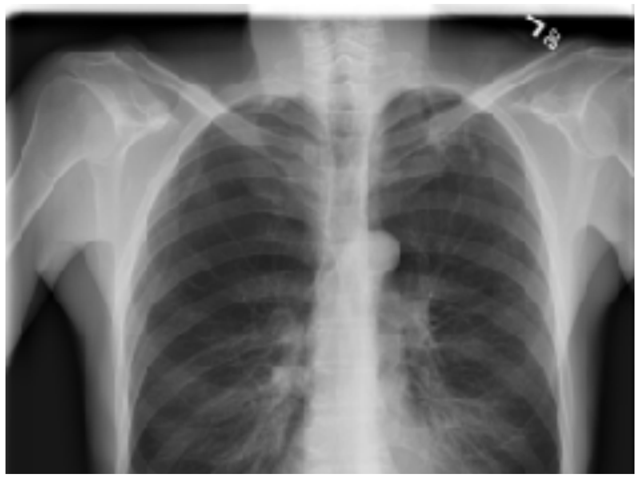


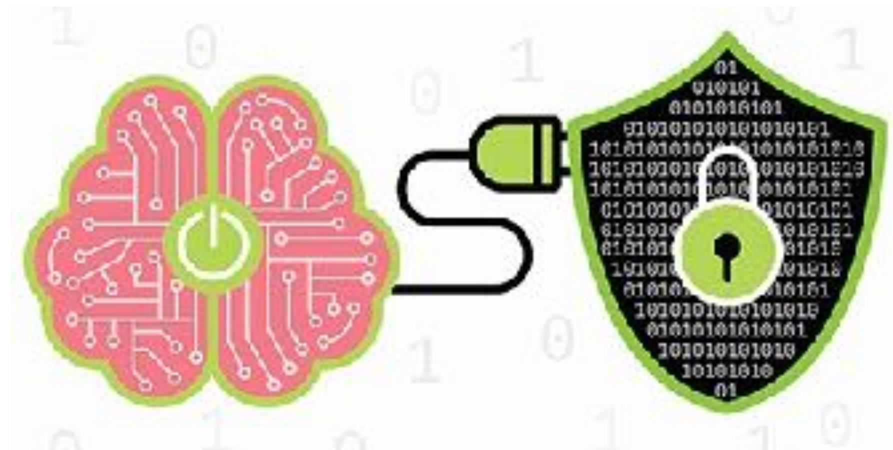
- Example: real-life data analysis of microcredit to determine if it helps alleviate poverty. We show: can remove <u>one</u> data point (out of 16,500) to change sign of the effect

[https://stanfordmlgroup.github.io/projects/chexnet/, https://www.ibmbigdatahub.com/blog/cyber-security-powered-ai-and-machine-learning, https://www.fairwinds.org/inside-fairwinds/fraud-protection-center/report-fraud.html]

When can I trust my data analysis?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions



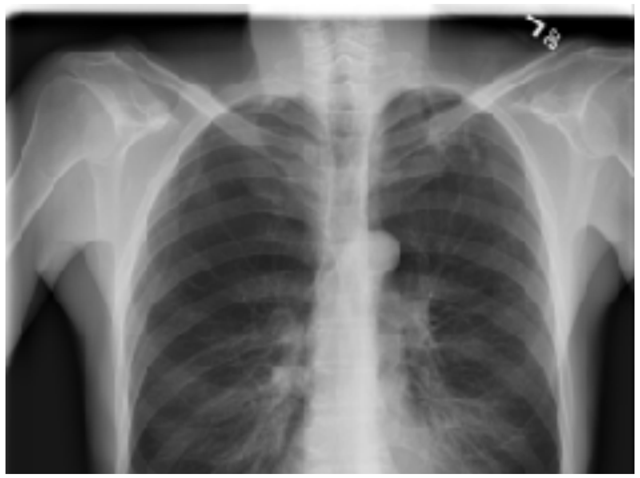
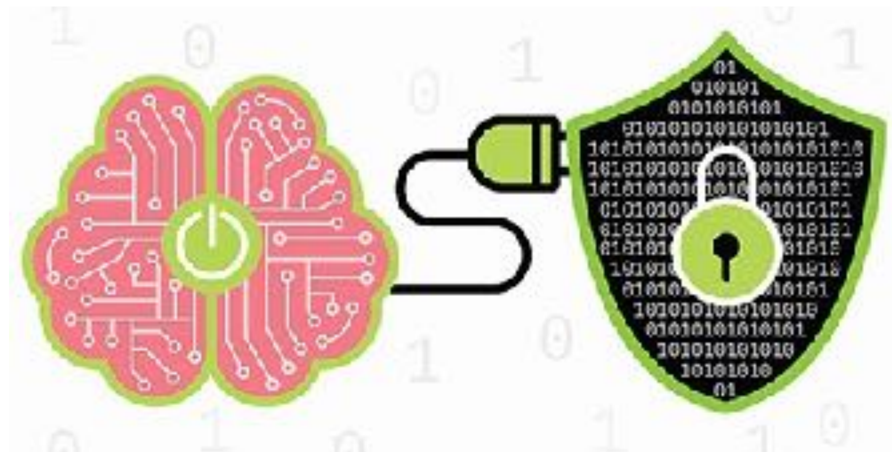


- Example: real-life data analysis of microcredit to determine if it helps alleviate poverty. We show: can remove <u>one</u> data point (out of 16,500) to change sign of the effect
- We find: Bayes & more model levels don't stop sensitivity

[https://stanfordmlgroup.github.io/projects/chexnet/, https://www.ibmbigdatahub.com/blog/cyber-security-powered-ai-and-machine-learning, https://www.fairwinds.org/inside-fairwinds/fraud-protection-center/report-fraud.html]

When can I trust my data analysis?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions

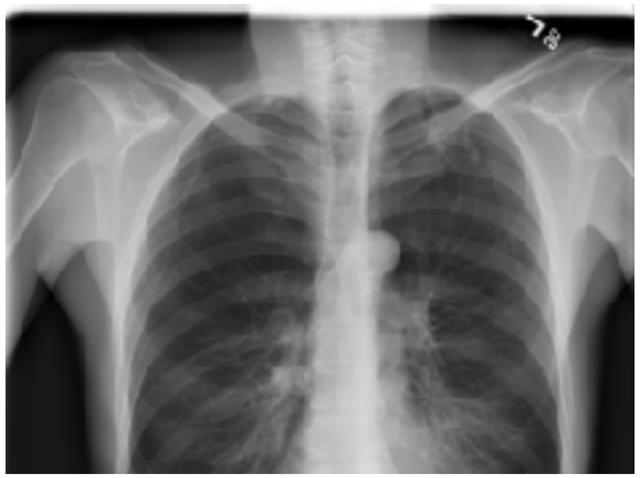



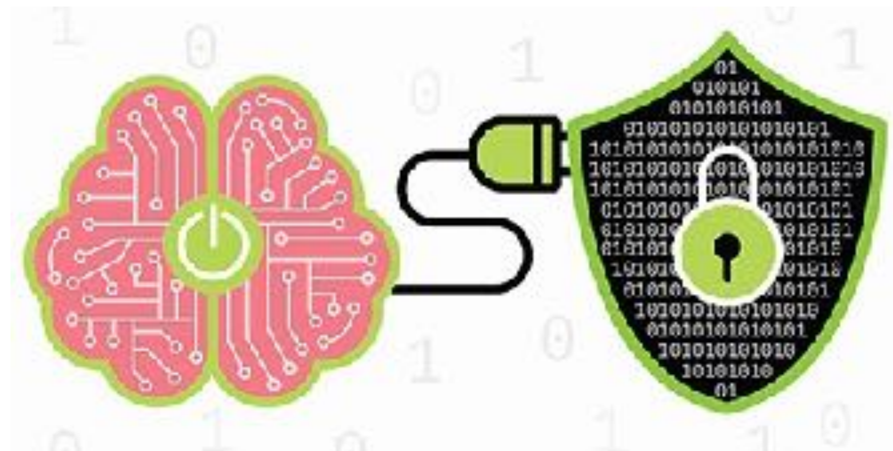

- Example: real-life data analysis of microcredit to determine if it helps alleviate poverty. We show: can remove <u>one</u> data point (out of 16,500) to change sign of the effect
- We find: Bayes & more model levels don't stop sensitivity
- How can we find this sensitivity in practice?

[https://stanfordmlgroup.github.io/projects/chexnet/, https://www.ibmbigdatahub.com/blog/cyber-security-powered-ai-and-machine-learning, https://www.fairwinds.org/inside-fairwinds/fraud-protection-center/report-fraud.html]

When can I trust my data analysis?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions



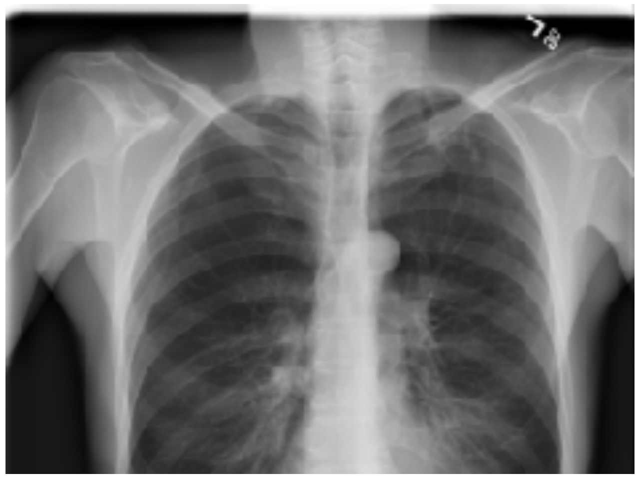


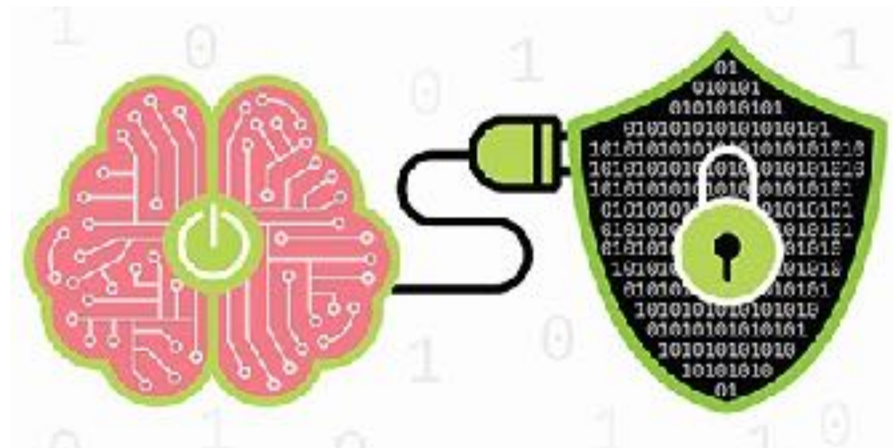
- Example: real-life data analysis of microcredit to determine if it helps alleviate poverty. We show: can remove <u>one</u> data point (out of 16,500) to change sign of the effect
- We find: Bayes & more model levels don’t stop sensitivity
- How can we find this sensitivity in practice?
- **Challenge**: Too expensive to check every data subset

[https://stanfordmlgroup.github.io/projects/chexnet/, https://www.ibmbigdatahub.com/blog/cyber-security-powered-ai-and-machine-learning, https://www.fairwinds.org/inside-fairwinds/fraud-protection-center/report-fraud.html]

When can I trust my data analysis?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions





- Example: real-life data analysis of microcredit to determine if it helps alleviate poverty. We show: can remove <u>one</u> data point (out of 16,500) to change sign of the effect
- We find: Bayes & more model levels don't stop sensitivity
- How can we find this sensitivity in practice?
- **Challenge**: Too expensive to check every data subset
- **Our Solution**: a fast, automated, accurate *approximation*

[https://stanfordmlgroup.github.io/projects/chexnet/, https://www.ibmbigdatahub.com/blog/cyber-security-powered-ai-and-machine-learning, https://www.fairwinds.org/inside-fairwinds/fraud-protection-center/report-fraud.html]

Why care about dropping data subsets?

- Why would you care? It varies by problem.

Why care about dropping data subsets?

- Why would you care? It varies by problem.
- Thinking without random noise can be helpful.

Why care about dropping data subsets?

- Why would you care? It varies by problem.
- Thinking without random noise can be helpful.
  - Suppose you have a farm, and want to know whether your average yield is greater than 170 bushels per acre.

Why care about dropping data subsets?

- Why would you care? It varies by problem.
- Thinking without random noise can be helpful.
  - Suppose you have a farm, and want to know whether your average yield is greater than 170 bushels per acre.
    - At harvest, you measure 200 bushels per acre.

Why care about dropping data subsets?

- Why would you care? It varies by problem.
- Thinking without random noise can be helpful.
  - Suppose you have a farm, and want to know whether your average yield is greater than 170 bushels per acre.
    - At harvest, you measure 200 bushels per acre.
    - Scenario one: If your yield is greater than 170 bushels per acre, you make a profit.
      - Don't care about sensitivity to small subsets

Why care about dropping data subsets?

- Why would you care? It varies by problem.
- Thinking without random noise can be helpful.
    - Suppose you have a farm, and want to know whether your average yield is greater than 170 bushels per acre.
        - At harvest, you measure 200 bushels per acre.
        - Scenario one: If your yield is greater than 170 bushels per acre, you make a profit.
            - Don't care about sensitivity to small subsets
        - Scenario two: You want to recommend your farming methods to a friend across the valley.
            - Might care about sensitivity to small subsets

Why care about dropping data subsets?

Why care about dropping data subsets?

- Examples of specific reasons to worry about sensitivity to dropping small subsets:

Why care about dropping data subsets?

- Examples of specific reasons to worry about sensitivity to dropping small subsets:
  - Policy population different from analyzed population

Why care about dropping data subsets?

- Examples of specific reasons to worry about sensitivity to dropping small subsets:
  - Policy population different from analyzed population
  - Small fractions of data often missing not-at-random

Why care about dropping data subsets?

- Examples of specific reasons to worry about sensitivity to dropping small subsets:
  - Policy population different from analyzed population
  - Small fractions of data often missing not-at-random
  - Report a convenient proxy (e.g. mean)

Why care about dropping data subsets?

- Examples of specific reasons to worry about sensitivity to dropping small subsets:
  - Policy population different from analyzed population
  - Small fractions of data often missing not-at-random
  - Report a convenient proxy (e.g. mean)
  - Models are misspecified

Why care about dropping data subsets?

- Examples of specific reasons to worry about sensitivity to dropping small subsets:
  - Policy population different from analyzed population
  - Small fractions of data often missing not-at-random
  - Report a convenient proxy (e.g. mean)
  - Models are misspecified

- Our paper focuses on experiments from economics (in part due to fantastic reproducibility!)

Why care about dropping data subsets?

- Examples of specific reasons to worry about sensitivity to dropping small subsets:
  - Policy population different from analyzed population
  - Small fractions of data often missing not-at-random
  - Report a convenient proxy (e.g. mean)
  - Models are misspecified

- Our paper focuses on experiments from economics (in part due to fantastic reproducibility!)
  - But these concerns and our techniques are much more general

Why do we need an approximation?

Why do we need an approximation?

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico.

Why do we need an approximation?

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico.
    - The study included ~16k households.

Why do we need an approximation?

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico.
  - The study included ~16k households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16k).

Why do we need an approximation?

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico.
  - The study included ~16k households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16k).
  - No matter how fast you fit, brute force is impossible.

Why do we need an approximation?

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico.
  - The study included ~16k households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16k).
  - No matter how fast you fit, brute force is impossible.
- Our approximation:
  1. Represents data dropping using data reweighting, then
  2. Forms a Taylor series approximation.

Why do we need an approximation?

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico.
  - The study included ~16k households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16k).
  - No matter how fast you fit, brute force is impossible.
- Our approximation:
  1. Represents data dropping using data reweighting, then
  2. Forms a Taylor series approximation.
  - Comes with non-stochastic accuracy guarantees.

Why do we need an approximation?

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico.
  - The study included ~16k households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16k).
  - No matter how fast you fit, brute force is impossible.
- Our approximation:
  1. Represents data dropping using data reweighting, then
  2. Forms a Taylor series approximation.
  - Comes with non-stochastic accuracy guarantees.
  - Can be checked by re-fitting only once.

Why do we need an approximation?

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico.
    - The study included ~16k households.
    - There are ~10^{53} subsets of size 16 (0.1% of 16k).
    - No matter how fast you fit, brute force is impossible.
- Our approximation:
    1. Represents data dropping using data reweighting, then
    2. Forms a Taylor series approximation.
    - Comes with non-stochastic accuracy guarantees.
    - Can be checked by re-fitting only once.
    - Works for variational Bayes, MAP, MLE, OLS, etc.
        - All minimizers of smooth empirical loss!

Experimental results

- Are Bayesian analyses necessarily more robust?

Experimental results

- Are Bayesian analyses necessarily more robust?
- We study a VB approximation to Meager (2020)

Experimental results

- Are Bayesian analyses necessarily more robust?
- We study a VB approximation to Meager (2020)
  - Hierarchical model on multiple microcredit studies

Experimental results

- Are Bayesian analyses necessarily more robust?
- We study a VB approximation to Meager (2020)
    - Hierarchical model on multiple microcredit studies
    - Likelihood and priors carefully chosen

Experimental results

- Are Bayesian analyses necessarily more robust?
- We study a VB approximation to Meager (2020)
  - Hierarchical model on multiple microcredit studies
  - Likelihood and priors carefully chosen
  - VB matches MCMC output from Stan

Experimental results

- Are Bayesian analyses necessarily more robust?
- We study a VB approximation to Meager (2020)
  - Hierarchical model on multiple microcredit studies
  - Likelihood and priors carefully chosen
  - VB matches MCMC output from Stan
    - …but only when using linear response covariances (Giordano et al. (2018))

Experimental results

- Are Bayesian analyses necessarily more robust?
- We study a VB approximation to Meager (2020)
  - Hierarchical model on multiple microcredit studies
  - Likelihood and priors carefully chosen
  - VB matches MCMC output from Stan
    - …but only when using linear response covariances (Giordano et al. (2018))
  - We find that dropping < 0.1% of data changes the sign of the posterior expected average effect of microcredit
    - Similar sensitivity to the OLS analyses!

Experimental results

Experimental results

- Beyond Bayes:
  - Oregon Medicaid study (Finkelstein et al 2012): winners of a lottery could sign up for Medicaid

Experimental results

- Beyond Bayes:
  - Oregon Medicaid study (Finkelstein et al 2012): winners of a lottery could sign up for Medicaid
    - $p < 0.01$ for a positive effect of lottery on health

Experimental results

- Beyond Bayes:
  - Oregon Medicaid study (Finkelstein et al 2012): winners of a lottery could sign up for Medicaid
    - $p < 0.01$ for a positive effect of lottery on health
    - We find: dropping 11 points (0.05%) of >21,000 data points changes statistical significance

Experimental results

- Beyond Bayes:
  - Oregon Medicaid study (Finkelstein et al 2012): winners of a lottery could sign up for Medicaid
    - $p < 0.01$ for a positive effect of lottery on health
    - We find: dropping 11 points (0.05%) of >21,000 data points changes statistical significance
  - More cash transfers (Angelucci and De Giorgi 2009)
    - Effect on poor households is robust!

Experimental results

- Beyond Bayes:
  - Oregon Medicaid study (Finkelstein et al 2012): winners of a lottery could sign up for Medicaid
    - $p < 0.01$ for a positive effect of lottery on health
    - We find: dropping 11 points (0.05%) of >21,000 data points changes statistical significance
  - More cash transfers (Angelucci and De Giorgi 2009)
    - Effect on poor households is robust!

- We show that, in general, sensitivity to dropping small data subsets is:

Experimental results

- Beyond Bayes:
    - Oregon Medicaid study (Finkelstein et al 2012): winners of a lottery could sign up for Medicaid
        - $p < 0.01$ for a positive effect of lottery on health
        - We find: dropping 11 points (0.05%) of >21,000 data points changes statistical significance
    - More cash transfers (Angelucci and De Giorgi 2009)
        - Effect on poor households is robust!

- We show that, in general, sensitivity to dropping small data subsets is:
    - Not primarily driven by misspecification, small sample sizes, or gross outliers

Experimental results

- Beyond Bayes:
  - Oregon Medicaid study (Finkelstein et al 2012): winners of a lottery could sign up for Medicaid
    - $p < 0.01$ for a positive effect of lottery on health
    - We find: dropping 11 points (0.05%) of >21,000 data points changes statistical significance
  - More cash transfers (Angelucci and De Giorgi 2009)
    - Effect on poor households is robust!

- We show that, in general, sensitivity to dropping small data subsets is:
    - Not primarily driven by misspecification, small sample sizes, or gross outliers
    - Is primarily driven by a low “signal to noise ratio”

Conclusions

- We develop a fast way to check if there is a very small fraction of data you can drop to change conclusions
  - **"An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?" https://arxiv.org/abs/2011.14999**

Conclusions

- We develop a fast way to check if there is a very small fraction of data you can drop to change conclusions
  - **"An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?" https://arxiv.org/abs/2011.14999**
- See also:
  - Giordano, R, Broderick, T, and Jordan, MI. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 2018.
  - Giordano, R*, Liu, R*, Jordan, MI, and Broderick, T. (*joint first authorship) Evaluating Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics. https://arxiv.org/abs/2107.03584

Conclusions

- We develop a fast way to check if there is a very small fraction of data you can drop to change conclusions
  - **"An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?" https://arxiv.org/abs/2011.14999**
- See also:
  - Giordano, R, Broderick, T, and Jordan, MI. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 2018.
  - Giordano, R*, Liu, R*, Jordan, MI, and Broderick, T. (*joint first authorship) Evaluating Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics. https://arxiv.org/abs/2107.03584
- Reproducibility is a prerequisite to run our check

Conclusions

- We develop a fast way to check if there is a very small fraction of data you can drop to change conclusions
  - **"An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?" https://arxiv.org/abs/2011.14999**
- See also:
  - Giordano, R, Broderick, T, and Jordan, MI. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 2018.
  - Giordano, R*, Liu, R*, Jordan, MI, and Broderick, T. (*joint first authorship) Evaluating Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics. https://arxiv.org/abs/2107.03584
- Reproducibility is a prerequisite to run our check
  - "Transparency and Reproducibility in Artificial Intelligence," *Nature Matters Arising*, 2020.

Conclusions

- We develop a fast way to check if there is a very small fraction of data you can drop to change conclusions
  - **"An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?" https://arxiv.org/abs/2011.14999**
- See also:
  - Giordano, R, Broderick, T, and Jordan, MI. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 2018.
  - Giordano, R*, Liu, R*, Jordan, MI, and Broderick, T. (*joint first authorship) Evaluating Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics. https://arxiv.org/abs/2107.03584
- Reproducibility is a prerequisite to run our check
  - "Transparency and Reproducibility in Artificial Intelligence," *Nature Matters Arising*, 2020.
  - "Toward a Taxonomy of Trust for Probabilistic Machine Learning" https://arxiv.org/abs/2112.03270